Homework 6

Deepika Dilip

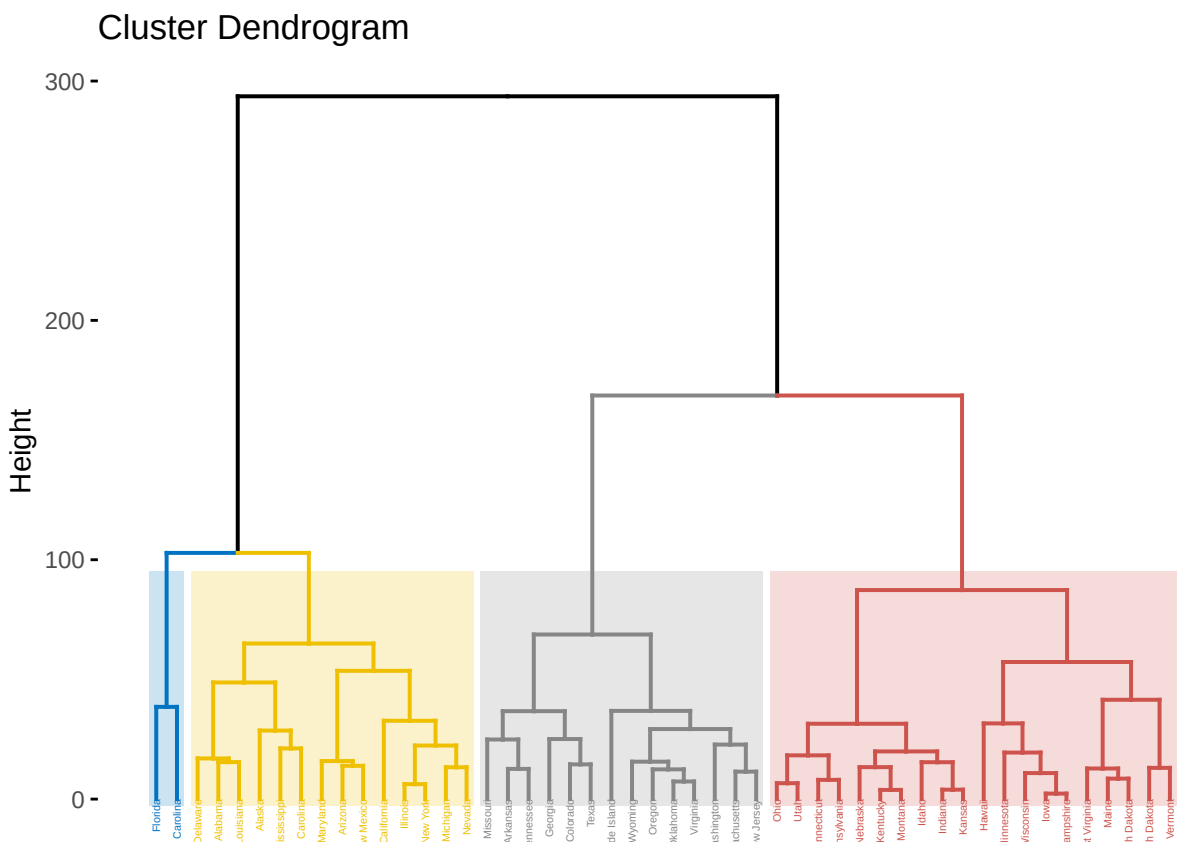
Cluster analysis

Part A

```
#Querying Dataset
USA <- USArrests

#USA <- scale(USA)


#Hierarchal Clustering
hc.complete <- hclust(dist(USA), method = "complete")


summary(hc.complete)
```

```
##             Length Class  Mode
## merge       98     -none- numeric
## height      49     -none- numeric
## order       50     -none- numeric
## labels      50     -none- character
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
fviz_dend(hc.complete,
          k = 4,
          cex = 0.3,
          palette = "jco",
          color_labels_by_k = TRUE,
          #type = c("circular"),
          rect = TRUE,
          rect_fill = TRUE,
          rect_border = "jco",
          labels_track_height = 2.5)
```



Part B

```
ind4.complete <- cutree(hc.complete, 3)
```

List of States in first cluster

```
USA[ind4.complete == 1,] %>% print()
```

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Arizona           8.1     294       80 31.0
## California        9.0     276       91 40.6
## Delaware          5.9     238       72 15.8
## Florida          15.4     335       80 31.9
## Illinois         10.4     249       83 24.0
## Louisiana        15.4     249       66 22.2
## Maryland         11.3     300       67 27.8
## Michigan         12.1     255       74 35.1
## Mississippi      16.1     259       44 17.1
## Nevada           12.2     252       81 46.0
## New Mexico       11.4     285       70 32.1
## New York         11.1     254       86 26.1
## North Carolina   13.0     337       45 16.1
```

```
## South Carolina   14.4     279       48 22.5
```

List of States in second cluster

```
USA[ind4.complete == 2,] %>% print()
```

```
##               Murder Assault UrbanPop Rape
## Arkansas         8.8     190       50 19.5
## Colorado         7.9     204       78 38.7
## Georgia         17.4     211       60 25.8
## Massachusetts    4.4     149       85 16.3
## Missouri         9.0     178       70 28.2
## New Jersey       7.4     159       89 18.8
## Oklahoma         6.6     151       68 20.0
## Oregon           4.9     159       67 29.3
## Rhode Island     3.4     174       87  8.3
## Tennessee       13.2     188       59 26.9
## Texas           12.7     201       80 25.5
## Virginia         8.5     156       63 20.7
## Washington       4.0     145       73 26.2
## Wyoming          6.8     161       60 15.6
```

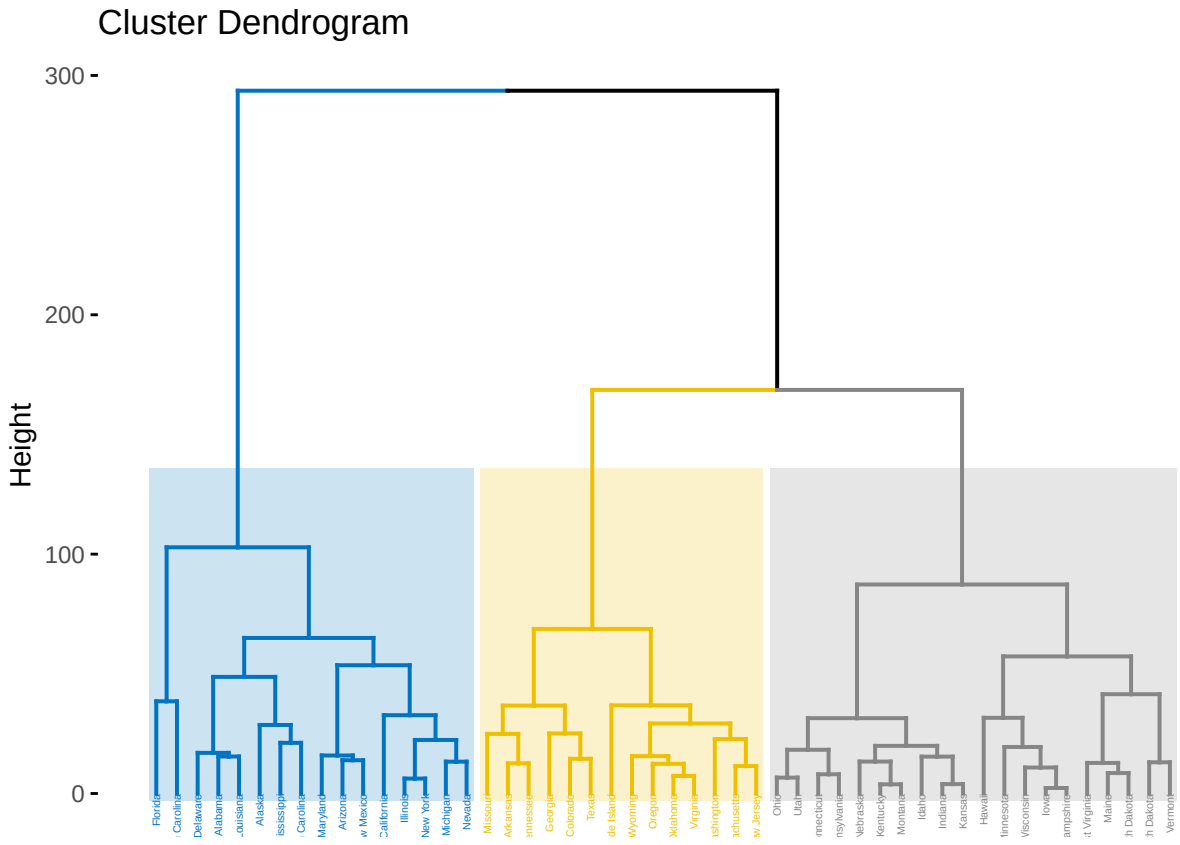
List of States in third cluster

```
USA[ind4.complete == 3,] %>% print()
```

```
##               Murder Assault UrbanPop Rape
## Connecticut      3.3     110       77 11.1
## Hawaii           5.3      46       83 20.2
## Idaho            2.6     120       54 14.2
## Indiana          7.2     113       65 21.0
## Iowa             2.2      56       57 11.3
## Kansas           6.0     115       66 18.0
## Kentucky         9.7     109       52 16.3
## Maine            2.1      83       51  7.8
## Minnesota        2.7      72       66 14.9
## Montana          6.0     109       53 16.4
## Nebraska         4.3     102       62 16.5
## New Hampshire    2.1      57       56  9.5
## North Dakota     0.8      45       44  7.3
## Ohio             7.3     120       75 21.4
## Pennsylvania     6.3     106       72 14.9
## South Dakota     3.8      86       45 12.8
## Utah             3.2     120       80 22.9
## Vermont          2.2      48       32 11.2
## West Virginia    5.7      81       39  9.3
## Wisconsin        2.6      53       66 10.8
```

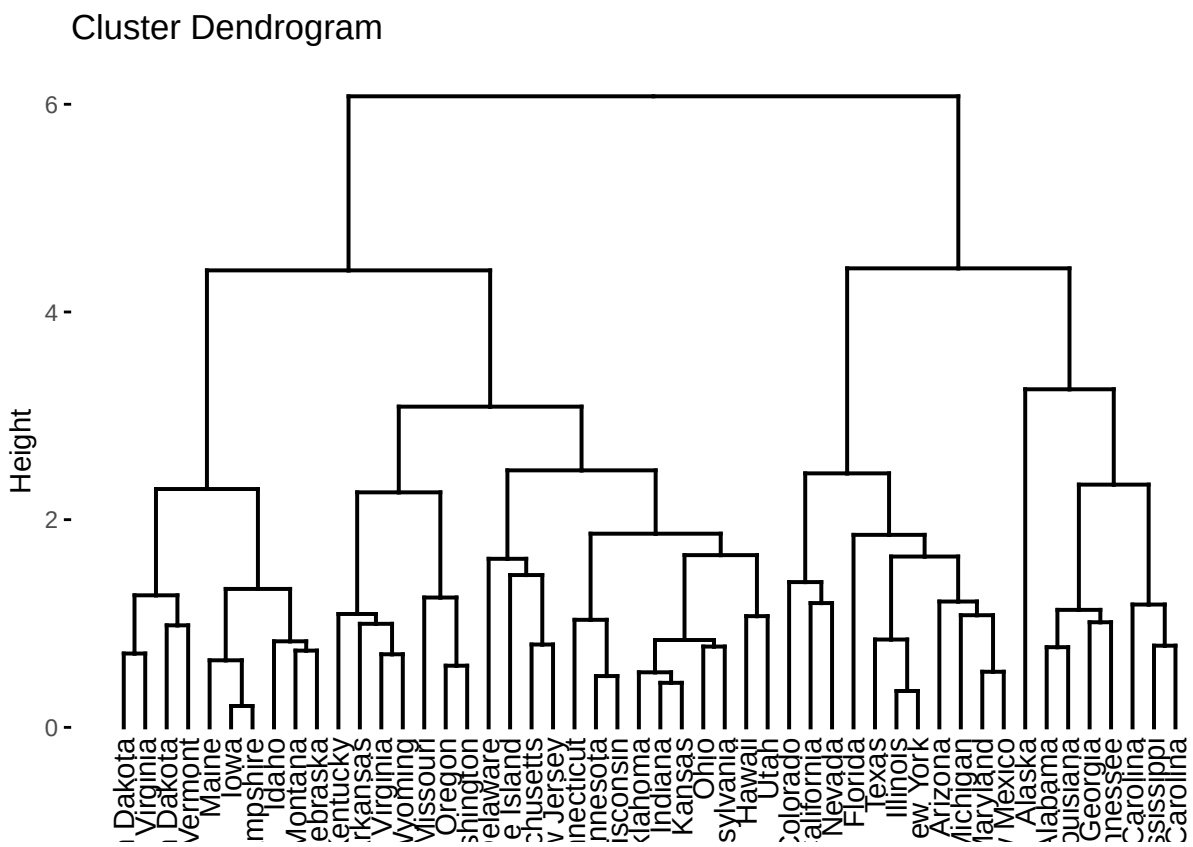
Dendogram

```
fviz_dend(hc.complete,
          k = 3,
          cex = 0.3,
          palette = "jco",
          color_labels_by_k = TRUE,
          #type = c("circular"),
          rect = TRUE,
          rect_fill = TRUE,
          rect_border = "jco",
          labels_track_height = 2.5)
```



Part C

```
USA_scaled <- scale(USA)
scaling <- function(x) (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)

hc.complete.scaled <- hclust(dist(USA_scaled), method = "complete")

fviz_dend(hc.complete.scaled ) %>% plot()
```



Part D:

The goal of the algorithm behind hierarchical clustering is to minimize dissimilarlity between clusters. In this case, scaling resulted in more clusters along with additional tree height. The method we use depends on the nature of the data and the question we are trying to answer. In my opinion, I prefer the more parsimonious model (i.e. pre-scaling). In this case (in my opinion), scaling wasn't as necessary, given that most of the variables were proportions (per 100,000) and another was a percentage.

Problem 2: PCA

```
img <- readJPEG('tiger.jpg')

dim(img)
```

```
## [1] 884 884   3
```

```
r <- img[,,1]
g <- img[,,2]
b <- img[,,3]

img.r.pca <- prcomp(r, center = FALSE)
img.g.pca <- prcomp(g, center = FALSE)
img.b.pca <- prcomp(b, center = FALSE)
```

```r
rgb.pca <- list(img.r.pca, img.g.pca, img.b.pca)

# Approximate X with XV_kV_k^T
compress <- function(pr, k)
{
  compressed.img <- pr$x[,1:k] %*% t(pr$rotation[,1:k])
  compressed.img
}

# Using first 20 PCs
pca20 <- sapply(rgb.pca, compress, k = 20, simplify = "array")

writeJPEG(pca20, "tiger20.jpeg")


# 50 PCS
pca50 <- sapply(rgb.pca, compress, k = 50, simplify = "array")

writeJPEG(pca50, "tiger50.jpeg")



# 100 PCS
pca100 <- sapply(rgb.pca, compress, k = 100, simplify = "array")

writeJPEG(pca100, "tiger100.jpeg")


#200 PCS

pca200 <- sapply(rgb.pca, compress, k = 200, simplify = "array")

writeJPEG(pca200, "tiger200.jpeg")
```

Interpretation:

As the number of principal components increases, so does the clarity of the image. However (in my artistic opinion), the clarity of the image did increase dramatically from 20 to 50 principal components and from 50 to 100 principal components but not so much from 100 to 200 principal components. We need to consider this in real-life applications, especially when figuring out the cut-off for facial recognition software, which has social ramifications.

Figure 1: PCA20

Figure 2: PCA50

Figure 3: PCA100

Figure 4: PCA200